BIOINFORMATICS APPLICATIONS NOTE
Vol. 21 no. 8 2005, pages 1699–1700
doi:10.1093/bioinformatics/bti194

Genome analysis

HESAS: HERVs Expression and Structure Analysis System

Tae-Hyung Kim[1], Yeo-Jin Jeon[1], Woo-Yeon Kim[1] and Heui-Soo Kim[1,2,*]

[1]PBBRC, Interdisciplinary Research Program of Bioinformatics and [2]Division of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609–735, Korea

Received on October 21, 2004; revised on November 11, 2004; accepted on November 26, 2004
Advance Access publication December 10, 2004

ABSTRACT

Summary: HESAS (HERVs Expression and Structure Analysis System) database was developed to understand the human endogenous retroviruses (HERVs) that have an effect on the expression of human functional genes. The database products are generated by the exon-based expressed sequence tag clustering and reconstructing of partial HERV structures that result from various mutations during primate evolution. The expression types were classified according to the existence of splicing, transcriptional start and polyadenylation signal sites. The database currently contains HERV information on 26 981 human genes of exon–intron structure. The HERV elements were inserted into 17 317 of these genes and linked to expression with 898 genes.

Availability: http://www.primate.or.kr/HESAS

Contact: khs307@pusan.ac.kr

1 INTRODUCTION

The human genome contains various endogenous retroviruses (HERVs) that represent the footprints of ancient germ-cell infections (Lower *et al*., 1996). These HERVs and other long terminal repeat (LTR)-like elements account for ∼8% of the human genome. Most of the HERVs are no longer able to code for functional protein, owing to multiple stop codons, insertions, deletions and frame shifts. However, the presence of LTR-promoters from some HERV families are increasing their expression in human placenta and several cancer cell lines (Mi *et al*., 2000). Recently, these retroelements have gained the evolutionary potential role to enhance the coding capacity and regulatory versatility of the genome without compromising its integrity (Sorek *et al*., 2002). Database for HERV elements was used for searches of individual HERV families (Paces *et al*., 2002). Over the past decade, a considerable number of studies have been conducted on the capacity to modify the expression of neighboring genes (Jordan *et al*., 2003). These reports are focused mainly on strong LTR-derived promoters in the specific tissues, including data for alternative splicing and primary polyadenylation signal (Mager *et al*., 1999). However, there has been no study concerning systematic analyses of the creation of various transcripts caused by HERV elements within genomic sequences of human genes. Here, we characterized HERV positions, LTR-truncated constructs and HERV ORFs within the human genes. In addition, we present a large number of HERVs linked to genes that are expressed in normal tissues and pathology tissues using an electronic mapping method.

2 ANALYSIS PIPELINE AND DATABASE CONSTRUCTION

2.1 Data sources

Intron and both of the 5 kb regions, upstream- and downstream-containing genes for the HESAS (HERVs Expression and Structure Analysis System), were obtained from GoldenPath, which is based on gene information of NCBI Build 34.3. Intron/exon structures forming various HERV-related transcripts in genes are obtained by alignment of RefSeq mRNAs as counterparts of the human genes (Pruitt *et al*., 2003). HERV elements in genes were identified by RepeatMasker (http://repeatmasker.genome.washington.edu), an embedded MaskerAid (Bedell *et al*., 2000) with 352 ERV consensus sequences from the Repbase Update (Jurka, 2000). A HERV's locus on the genome, its position in the consensus sequence, direction, subfamily name and Smith–Waterman score were derived from RepeatMasker's outputs. The expressed sequence tag (EST) sequences were derived from NCBI's dbEST database that contains 8209 cDNA libraries (Boguski *et al*., 1993). The useful EST information for tissues and pathology types was obtained from the eVOC ontology, a set of controlled vocabularies for unifying gene expression data (Kelso *et al*., 2003).

2.2 EST clustering

For EST clustering with strict supervisor to detect HERV-related expression patterns from a given set of ESTs, intron/exon structures resulting from the mapping of RefSeq mRNA by sim4 (Florea *et al*., 1998) were collected. We set the criterion that at least one side of the two-end boundary of an exon region had to overlap with an aligned EST with an identity >97% in order to eliminate EST data contaminated by genomic sequences. To obtain expressed transcripts with HERVs only, non-ERV repeat sequences such as LINE, SINE, MIR and simple repeat elements within all genes were masked by RepeatMasker before performing EST clustering.

2.3 Identification of HERV genomic structure

Consensus domain libraries for scanning HERVs were newly constructed by comparing highly conserved residues as potential coding regions (*gag*, *pro*, *RT*, *RNaseH*, *IN* and *env*) within internal-HERV with traditional viral genes of Pfam (Bateman *et al*., 2004) using HMMER. Thus, both libraries, of the consensus domain and the original Repbase, were simultaneously used to identify the HERV elements within whole gene region. We found that 17 317 of the 26 981 human genes were inserted by HERVs in the inner or neighborhood part of their gene region. The genomic structure of HERV

[*]To whom correspondence should be addressed.

elements was defined as LTR-HERV-LTR including the boundary of the coding region. These were divided into four types (complete, 5′ truncate, 3′ truncate and 5′–3′ truncate) according to LTR truncations, and solitary LTRs that were detected by calculating the RepeatMasker output in reconstructing the HERV structure of the non-fragmental state before accumulating mutation (Kim *et al*., 2004).

2.4 Analysis of the types of HERV expression within human genes

The genomic locus of the HERVs and exon information on the splicing structure of mRNA/EST were calculated from their positions within the genes. The HERV expression was interpreted as an overlap relationship of HERV and mRNA/EST within the genes. All the overlapping states are determined by fully or partially depending on inclusion relationship. The types of HERV expression were classified according to the position of HERVs within the genes and information on the splicing structure of mRNA/EST.

3 ACCESS AND VISUALIZATION

The HESAS can be searched in two major modes. First, users can query to detect the types of expressed transcripts with HERVs and the genomic structure of HERVs by selecting the chromosome number and clicking on ‘HERV expression’. Second, it is possible for users to search the existence of HERV expression by selection of specific or interesting genes, and EST category of tissues or pathology. Result pages are listed in the tabular format to represent the evidence and information of expressed HERV events within genes. Using Java applet, we also developed a viewer of sufficient power to analyze types of HERV expression. This viewer shows a transcript of each expressed HERV that is represented by the exon/intron splicing structure of mRNAs/ESTs, as well as merging HERV elements as evidence of expressed HERV elements. Moreover, the event of expressed HERV provides a visual presentation of various highlighted transcript structures, alternative promoter, splicing and polyadenylation signal. Viewers are allowed a facilitative function to zoom in and out of any region of a gene to show a sophisticated image. The 2 bp flanking region of the exon boundary is compared with the given canonical splicing site (AG-GT), and then it is represented by symbolic characters.

REFERENCES

Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.

Bedell,J.A., Korf,I. and Gish,W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**, 1040–1041.

Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for “expressed sequence tags”. *Nat. Genet*., **4**, 332–333.

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.

Jordan,I.K., Rogozin,I.B., Glazko,G.V. and Koonin,E.V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet*., **19**, 68–72.

Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*., **16**, 418–420.

Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien,S., Smedley,D., Otgaar,D., Greyling,G., Jongeneel,C.V., McCarthy,M.I. *et al*. (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res*., **13**, 1222–1230.

Kim,T.H., Jeon,Y.J., Yi,J.M., Kim,D.S., Huh,J.W., Hur,C.G. and Kim,H.S. (2004) The distribution and expression of HERV families in the human genome. *Mol. Cells*, **18**, 87–93.

Lower,R., Lower,J. and Kurth,R. (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc. Natl Acad. Sci. USA*, **93**, 5177–5184.

Mager,D.L., Hunter,D.G., Schertzer,M. and Freeman,J.D. (1999) Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (*HHLA2* and *HHLA3*). *Genomics*, **59**, 255–263.

Mi,S., Lee,X., Li,X., Veldman,G.M., Finnerty,H., Racie,L., LaVallie,E., Tang,X.Y., Edouard,P., Howes,S. *et al*. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, **403**, 785–789.

Paces,J., Pavlicek,A. and Paces,V. (2002) HERVd: database of human endogenous retroviruses. *Nucleic Acids Res*., **30**, 205–206.

Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI reference sequence project: update and current status. *Nucleic Acids Res*., **31**, 34–37.

Sorek,R., Ast,G. and Graur,D. (2002) Alu-containing exons are alternatively spliced. *Genome Res*., **12**, 1060–1067.